The scalability of an application is the measure of the number of client requests it can simultaneously handle. When a hardware resource runs out and can no longer handle requests, it is counted as the limit of scalability. When this limit of the resource is reached, the application can no longer handle additional requests. To efficiently handle additional requests, administrators should scale the infrastructure by adding more resources such as RAM, CPU, storage, network devices, etc. Horizontal and vertical scaling are the two methods implemented by administrators for capacity planning.

Scalability is a crucial requirement of a cloud environment. You need to dynamically increase or decrease IT capacity or size to meet changing business IT requirements and manage unexpected traffic spikes. It will reduce latency and improve performance while preventing downtimes.

Scaling of two types

🡪Vertical scaling

🡪Horizontal Scaling

Horizontal Scaling

🡪Horizontal scaling is an approach of adding more devices to the infrastructure to increase the capacity and efficiently handle increasing traffic demands.

🡪Horizontal scaling is about expanding the capacity horizontally by adding extra servers. The load and processing power are shared among multiple servers within a system using a load balancer. It is also called scaling out.

Vertical Scaling

🡪Vertical scaling is a type of scalability wherein more computing and processing power is added to a machine to increase its performance. Also called scale-up, vertical scaling allows you to increase the machine’s capacity while maintaining resources within the same logical unit. The processor, memory, storage, and network capacity are increased in this approach.

|  |  |  |
| --- | --- | --- |
| **Vertical Vs Horizontal Scaling** | **Vertical scaling** | **Horizontal Scaling** |
| **Data** | Data is executed on a single node | Data is partitioned and executed on multiple nodes |
| **Data Management** | Easy to manage – share data reference | Complex task as there is no shared address space |
| **Downtime** | Downtime while upgrading the machine | No downtime |
| **Upper limit** | Limited by machine specifications | Not limited by machine specifications |
| **Cost** | Lower licensing fee | Higher licensing fee |